Submission in Response to NSF CI 2030 Request for Information

DATE AND TIME: 2017-04-05 21:12:09
REFERENCE NO: 309

Author Names & Affiliations

- Mark Miller - UC San Diego
- Maureen O'Leary - Stony Brook University

Contact Email Address (for NSF use only)

(Hidden)

Research Domain, discipline, and sub-discipline

Biology, Comparative biology, paleontology, evolution, stratigraphy, paleobiology, evolution and development, comparative genomics, computer science, databases, cyberinfrastructure

Title of Submission

Cyberinfrastrucutre for Assemblimng the Tree of Life

Abstract (maximum ~200 words).

Establishing the evolutionary history of life on earth (i.e., assembling the Tree of Life) has been identified by Science magazine in 2005 as one of the top 125 scientific challenges of the new millenium. Accomplishing this goal means acquiring extensive genotypic and phenotypic inventories of past and present life on earth, and analyzing these data to infer the Tree of Life. Both data acquisition and analysis require a robust set of cyberinfrastructure tools, including software environments, that 1) support the acquisition of DNA sequence data and phenotypic data, and 2) link data discovery with improved algorithmic tools for phylogenetic inference run on scalable computational resources. Inferring the Tree of Life will also require curated and sustainable databases that 1) integrate phenotypic data of many types, 2) link phenotypic data with genotypic data, and 3) support expansion of the tree through rapidly emerging metagenomics techniques. Multiple divisions at NSF (BIO, CISE, GEO) have sponsored the development of important parts of this pipeline but it remains a key research challenge to integrate and sustain these entities as a major informatics research that will impact US and international scientific endeavors.

Question 1 Research Challenge(s) (maximum ~1200 words): Describe current or emerging science or engineering research challenge(s), providing context in terms of recent research activities and standing questions in the field.

Assembling the Tree of Life (defining the phylogenetic relationships between all living and extinct species) was identified by Science magazine in 2005 as one of the 125 remaining grand challenges at the forefront of modern scientific research (1). Research that provides the basis for an all-inclusive Tree of Life will also bring discoveries that benefit every field in biology and many dimensions of the earth sciences by laying the groundwork for understanding how the phenotype and the genotype are integrated. Assembling the Tree of Life ultimately requires a complete phenotypic and genotypic inventory of extant and extinct species. With these data all information will be in place to conduct robust phylogeny reconstruction, and to examine how, through the process of development, the genotype gives rise to the phenotype and how such biological processes have evolved over the course of Deep Time. NSF's BIO, GEO and CISE Divisions have recognized the importance of this research challenge and have spearheaded the initial development of key cyber resources to support the mission. The next two decades will require further investments to harden and integrate these initial resources, as well as creating and integrating new resources as required by advances in technology.

The goal of creating full genotypic and phenotypic inventories is currently being driven rapidly forward by advances in DNA sequencing technology and by improved tools for characterizing phenotypes. It is essential to progress in building the Tree of Life that these data are readily discoverable and accessible in public databases. Inventories of macroscopic organisms are expanding rapidly, with sequencing data stimulating rapid discovery of new species and the placement of sequenced species in the Tree of Life. Teams of comparative biologists and paleontologists are integrating phenotype data into tree building in innovative ways that can be greatly expanded. These include collecting tens of thousands of observations and images in phylogenetic databases and analyzing genomic and phenomic data in simultaneous analysis that improves tree structure and provides new information about the ages of clades. Integration of data with stratigraphic data recording evolutionary events in Deep Time also allows us to discover the time of origin of major features such as cell structures, feathers, leaves, and bipedalism. Characterizing morphological traits enhances our understanding of the evolutionary process, allowing us to better understand the position of extinct species within the Tree of Life.

Strengthening our understanding of the link between genotype and phenotype across the Tree of Life is critical to realizing the benefits of our rapidly expanding genetic inventory of life on earth. While possessing a genomic inventory of all living things is useful, it is just a beginning. Discoveries that benefit society will come from characterizing the phenotype of these organisms. It is through characterizing phenotype that we make links between genomic sequences and trait evolution, seeding discoveries such as new metabolic pathways and biomaterials, or major patterns of macroevolution elucidating the interdependence of communities of organisms, and a better understanding of the ecology of the planet.

Establishing genotype:phenotype connections is inherently slower and more difficult than acquiring DNA sequence alone, and will require development of cyberinfrastructure tools that are tuned to making phenotypic data discoverable across all of the various phenotypic data types, from metabolic pathways to 3D photographic images. Because of the greater challenge currently required for phenotype discovery, it is especially important to provide access to any existing data, as well as to create new tools and technologies to speed the discovery of genotype:phenotype linkages. NSF has supported the development of key resources for acquiring, sharing and archiving data on phenotypes and their use in the Tree of Life such as Digimorph and MorphoBank. However, these initiatives, sponsored as awards to individual investigator laboratories have not had the resources to stand on par with the major public resource of Genbank. Creating the necessary database resource will require large investment and integration of existing resources, to facilitate analyses that combine both genomic and phenomic data. NSF also supported the creation of CIPRES as a premiere global resource for efficient tree building using both phenomic and genomic data. The CIPRES resource provides singular access to contemporary algorithms including those that incorporate stratigraphic data to estimate the times of major events in the history of life.

In addition to advances through integrating phenotypic and genotypic data, surveys of diverse environments with new metagenomic sequencing techniques are dramatically expanding the number of known microbial species. Recently, Hug et al. (Nature Microbiology1;16048) reported a previously unknown major phyla radiation near the root of the bacterial tree that was undetectable by classical culturing techniques. This new branch nearly doubles the size of the known bacterial tree. Ongoing research promises to expand the Archaea and Eukaryotes similarly, adding new phyla that have been previously been undetectable. With thousands of new draft microbial genomes being created from environmental samples, the alpha taxonomic processing, vetting, and placement of these new genomes is in itself a very significant logistical task that will require supporting cyberinfrastructure.

Question 2 Cyberinfrastructure Needed to Address the Research Challenge(s) (maximum ~1200 words): Describe any limitations or absence of existing cyberinfrastructure, and/or specific technical advancements in cyberinfrastructure (e.g. advanced computing, data infrastructure, software infrastructure, applications, networking, cybersecurity), that must be addressed to accomplish the identified research challenge(s).

Completing a genetic and phenotypic inventory of life on earth, and assembling those genomes into a coherent evolutionary tree requires cyberinfrastructure that supports (1) acquiring, discovering, sharing, and assembling phenomic and genomic data for analysis and storing it in ways that ensure reproducibility and (2) accessing community codes for data analyses on computational resources that scale to the large and growing needs of these analyses.

Specific tools that are needed for acquiring, discovering, sharing, and assembling phenomic data include:

1. Easy to access and stable formats for organizing and viewing large amounts of phenomic and genomic data during the construction of

the phylogenetic matrices. This includes not only web-based gene alignment and viewing software but also web applications and databases that store scientific observations on morphological homology that are supported by text and media (2D and 3D). This vision would expand both the contents and the capabilities of resources already built through NSF funding such as Digimorph and MorphoBank. And the creation of environments that integrate access to phenomic data with access to sequence data.

2. Cyberinfrastructure that supports the acquisition of phenotypic traits from digital images. Phenotypic image data are not as easily captured as data on genes and proteins, with the latter two types of data reducing to more readily manageable strings of discrete letters. Phenotype images present a more challenging information science problem for both data gathering and data searching, but one that can be met creatively. The challenge of capturing phenome images involves solving the problem of obtaining, storing and making a range of data types searchable: text descriptions of phenotypes in spreadsheet format, legacy and newly created media such as 2D and 3D images of phenotypes, and video and sound of physiological and behavioral processes, to name a few examples. NSF has supported creation of this type of cyberinfrastructure through projects such as MorphoBank, Digimorph, the PaleoBiology Database, and Encyclopedia of Life.

3. Sustainable, integrated database(s) of phenotypic traits. Currently, the NSF and other federal agencies have created a number of phenotype databases that are focused a specific clade or set of characters. Several databases exist for digital images alone. A substantial effort is required to create and annotate the data in each of the respective databases, but there is not a clear model for their integration or for their sustainability through time. It would be important to develop a strategic plan, not just to preserve the data in each of these databases, but to evaluate how the data in one might relevant to the others, and how that data might be discovered and presented in an integrated way. A resource that makes it possible to access, analyze and annotate, for example, all image data, through a single source could be a great benefit. A resource that provides integrates the many types of phenomic data from across the tree of life could be very useful.

4. Ideally the existing phenomic resources would be integrated into an end-to-end analysis resource. The imaging techniques developed under iDigBio (for example) would be seamlessly imported into a central database where those images can be accessed along with image data from the wide body of other image databases. Users can access images in the central data resource and use the collaboration and annotation technologies created by MorphoBank or Digimorph. To analyze the images collaboratively, and use these tools to create a matrix, and link their images to time data from the Paleobiology Database, and access to the necessary algorithmic tools to infer trees from an HPC provider such as CIPRES. Achieving this kind of integration can make the discovery process far more efficient, and would offer users a single source for access to phenotypic image data.

5. Access to community codes run on high performance computing clusters in a user-friendly environment. Access to parallel versions of important community codes installed on scalable clusters is essential, where users can choose access to command line tools and/or browser interfaces that do not require the user to invest in the overhead of learning to use an HPC machine from the command line, or Jupyter notebook interfaces, which support sharing and teaching of workflows. This could include investment in integrating existing data resources with software environments like Galaxy for DNA sequence assembly and CIPRES for tree inference.

6. Access to resources for metagenomic sequencing of environmental samples. With the explosive growth in new draft genomes for previously undescribed phyla, it is critical to have a central resource that can provide data access, a work area for preparing draft genomes, and HPC resources to help with their assembly. What is needed is a central repository that provides set of procedures for intake, curation, and publication of draft genomes sequences from unculturable organisms, and provides access to a curated data set that can be used to keep abreast of the current status of the Tree of Life, and can be used as a resource for developers of algorithmic tools for improved curation and assembly of environmental genomes.

Question 3 Other considerations (maximum ~1200 words, optional): Any other relevant aspects, such as organization, process, learning and workforce development, access, and sustainability, that need to be addressed; or any other issues that NSF should consider.

Process:
There is no doubt that appropriate cyberinfrastructure can advance the discovery process, by enabling new kinds of analyses, and by improving the productivity of individual researchers. In our view, enabling cyberinfrastructure has fundamental characteristics that are essential, and cyberinfrastructure that has these characteristics is both successful and rare.
1. Development is guided by the domain community that uses it, in collaboration with professional software developers and computer scientists. Creating cyberinfrastructure requires detailed knowledge of the problem to be addressed, and the finer details of how a domain

scientist works. Without that information, the project risks missing subtle but key requirements. Recommendation: CI projects should be led by someone actively doing research in the domain area who understands the workflow and algorithms of the community.
2. The resource must provide access to something the user cannot access in another way. It takes a significant amount of effort to learn to use a digital resource. If the user has already learned an alternative method, they will not adopt a new tool to do the same thing. Recommendation: CI projects must identify clear new functionality that will be created, and community interest in that function.
3. Skilled developers are required for sustainable software development. A project must be sufficiently reliable and professional to acquire a significant user base. In our experience this requires development by a skilled developer, as these individuals are far more likely to produce a sustainable CI package. There are many examples of digital resources that were no longer maintained after the graduate student who created it received their degree. Recommendation: CI projects must have a trained software developer on staff, the budget should be adequate for that. NSF should explore models for hiring, retaining, and sharing skilled developers of scientific software.
4. Establishing a digital resource takes a minimum of 5 years. It takes at least 3 years just to create a working digital resource. A three year funding cycle is not adequate to both create a resource and build a user base. Recommendation: CI projects should be provided with mechanisms for extended funding, conditional on meeting performance metrics. This could provide useful resources with an opportunity to gain traction in their community, yet avoid overcommitting resource to projects that are not successful, while along those that are to succeed.
Sustainability:
1. Models for how to sustain existing Cyberinfrastructure effectively are underdeveloped. Generally cyberinfrastructure is provided free of charge while a grant supports its creation and initial operations, but there is no mechanism for accepting payment, and once the project period is over, the cyberinfrastructure service ends. This produces infrastructure with a short lifetime, which is actually antithetical to the concept of infrastructure. Digital tools take years to create, and months to years to use effectively.

To address the problem of identifying funds to pay for infrastructure, one option would be to encourage investigators to include budget for supporting cyberinfrastructure in their proposals, or to offer credits which the PI can spend on the cyberinfrastructure of their choice. This would create a market-driven model for determining which infrastructure is sustained, and which is not. Credits would create a competitive marketplace, but would not have to be cash. This would simplify their use in a University environment. NIH Commons is currently conducting an interesting related experiment, where PIs are awarded credits which they can spend as they wish on approved cyberinfrastructure providers. This may be a model NSF could adopt.

Consent Statement